image-text pair like the Visual Question-Answering (VQA) task [46] or only an image like the image captioning task [47]. The output is the answer to the instruction conditioned on the input. The instruction template is flexible and subject to manual designs [21, 31, 33], as exemplified in Table 1. Note that the instruction samples can also be generalized to multi-round instructions, where the multimodal inputs are shared [21, 30, 31, 43].

Formally, a multimodal instruction sample can be denoted in a triplet form, *i.e.*, $(\mathcal{I}, \mathcal{M}, \mathcal{R})$, where $\mathcal{I}, \mathcal{M}, \mathcal{R}$ represent the instruction, the multimodal input, and the ground truth response, respectively. The MLLM predicts an answer given the instruction and the multimodal input:

$$\mathcal{A} = f(\mathcal{I}, \mathcal{M}; \theta) \tag{1}$$

Here, $\mathcal{A}$ denotes the predicted answer, and θ are the parameters of the model. The training objective is typically the original auto-regressive objective used to train the LLMs [21, 30, 32, 43], based on which the MLLM is forced to predict the next token of the response. The objective can be expressed as:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p(\mathcal{R}_i | \mathcal{I}, \mathcal{R}_{<i}; \theta) \tag{2}$$

where N is the length of the ground-truth response.

3.1.3 Modality Alignment

It is common to perform large-scale (compared to instruction-tuning) pre-training on paired data to encourage alignment between different modalities [25, 29, 35, 38], which is prior to the M-IT. The alignment datasets are typically image-text pairs [48–56] or Automatic Speech Recognition (ASR) [57–59] datasets, which all contain text. More specifically, the image-text pairs describe images in the form of natural language sentences, while the ASR datasets comprise transcriptions of speech. A common approach for alignment pre-training is to keep pre-trained modules (*e.g.* visual encoders and LLMs) frozen and train a learnable interface [21, 37, 38], which is illustrated in the following section.

3.1.4 Data

The collection of multimodal instruction-following data is a key to M-IT. The collection methods can be broadly categorized into benchmark adaptation, self-instruction [60], and hybrid composition. We illustrate these three methods sequentially.

Benchmark Adaptation Benchmark datasets are rich sources of high-quality data. Hence, abundant works [23–26, 28, 29, 32, 35] have utilized existing benchmark datasets to construct instruction-formatted datasets. Take the transformation of VQA datasets for an example, the original sample is an input-out pair where the input comprises an image and a natural language question, and the output is the textual answer to the question conditioned on the image. The input-output pairs of these datasets could naturally comprise the multimodal input and response of the instruction sample (see §3.1.2). The instructions, *i.e.*, the descriptions of the tasks, can either derive from manual design or from semi-automatic generation aided by GPT. Specifically, some works [13, 23, 25, 26, 36, 37] hand-craft a pool of candidate instructions and sample one of them during training. We offer an example of instruction templates for the VQA datasets as shown in Table 2. The other works manually design some seed instructions and use these instructions to prompt GPT to generate more [24, 31, 33].

Note that since the answers of existing VQA and caption datasets are usually concise, directly using these datasets for instruction tuning may limit the output length of MLLM. There are two common strategies to tackle this problem. The first one is to modify instructions. For example, ChatBridge [29] explicitly declares *short* and *brief* for short-answer data, as well as *a sentence* and *single sentence* for caption data. Similarly, InstructBLIP [23] inserts *short* and *briefly* into instruction templates for public datasets that inherently prefer short responses. The second one is to extend the length of existing answers [36]. For example, M^3IT [36] proposes to rephrase the original answer by prompting ChatGPT with the original question, answer, and context.

Self-Instruction Although existing benchmark datasets can contribute a rich source of data, they usually do not well meet human needs in real-world scenarios, such as multiple rounds of conversations. To tackle this issue, some works collect samples through self-instruction [60], which bootstraps LLMs to generate textual instruction-following data using a few hand-annotated samples. Specifically, some instruction-following samples are hand-crafted as seed examples, after which ChatGPT/GPT-4 is prompted to generate more instruction samples with the seed samples as guidance. LLaVA [21] extends the approach to the multimodal field by translating images into texts of captions and bounding boxes, and prompting GPT-4 to generate new data in the context of seed examples. In this way, an M-IT dataset is constructed, called LLaVA-Instruct-150k. Following this idea, subsequent works such as MiniGPT-4 [13], ChatBridge [29], GPT4Tools [34], and DetGPT [38] develop different M-IT datasets catering for different needs.

Hybrid Composition Apart from the M-IT data, language-only user-assistant conversation data can also be used to improve conversational proficiencies and instruction-following abilities [22, 31, 32, 35]. LaVIN directly constructs a mini-